


# Semantic Web Business Applications- A Scalability Model for the New Digital Economy

Sabina-Cristiana Necula

Alexandru Ioan Cuza University of Iasi

October 2012

# Semantic Web Business Applications- A Scalability Model for the New Digital Economy

SABINA-CRISTIANA NECULA
Department of Research
Alexandru Ioan Cuza University of Iasi
Carol I Blvd, no. 22, Iasi, 700505
ROMANIA
sabina.mihalache@gmail.com http://wwww.knowledgedecisionmaking.ro

*Abstract:* - Semantic web technologies are considered to be the next wave for web technologies related with rich internet web applications, content management, and document and information management. The most promising semantic web applications for business domain are considered to be the semantic web business portals which integrate diverse business information. Because semantic web applications are working with ontologies or data vocabularies there is a need to permanently assure the links between publicly available vocabularies on the web disposed at different addresses and diverse information which comes from different web sources. This means for semantic web business applications a scalability problem.
The present paper discusses the architecture of semantic web business application useful for assuring the scalability. We discuss the scalability problem in terms of data access and information retrieval. We conduct a series of experiments in order to test the scalability problems. Finally a so called scalability model is proposed.
The main contributions of the present paper consist in presenting the main problems that a semantic web business application presents in terms of scalability. We also contribute to the semantic web business applications field by presenting a framework to measure scalability.

*Key-Words:* - semantic web, scalability, model, new digital economy, RDF, OWL

## 1 Introduction

The digital economy is also sometimes called the Internet Economy, the New Economy, or Web Economy. In this new economy, digital networking and communication infrastructures provide a global platform over which people and organizations devise strategies, interact, communicate, collaborate and search for information [11].

The characteristics of interacting and communication are closely related to the aspects referring to the information exchange and, possibly, knowledge exchange in the form of best practices, case studies or other knowledge formalisms.

If the infrastructure is defined by the existing hardware capabilities and Internet technologies it cannot be said that the practical ways of collaborating are definitely defined. For the moment, users can exchange information over the Internet but any information provider has the interest of being the first in providing the meaningful information and the user has the interest to obtain rapidly and free the most relevant information.

A relatively new and very promising technology is the semantic web technology. The term and concept of semantic web is not so new, it is closely related to Artificial Intelligence but the practical application is very new, and it was initiated by the Linked Open Data project.

Since the whole Semantic Web community started to participate in the challenge that Linked Open Data proposed, diverse technologies appeared: tools for indexing information, tools to validate the structure of the files, tools to store information, standards to represent data, tools to access and query information, and tools to visualize information.

The main business activities or application fields in which semantic web technologies promise to offer great advantages are considered to be: E-Government, stock exchange, organizational portals and e-commerce/ advertising. It seems that any domain which deals with a lot of data coming from different sources or has a lot of meanings and it might be publicly available on the web is suitable for semantic web technology.

Apart from considering a relative new technology, semantic web proves to be suitable for the semantic requirements posed by business applications: E-government applications need data

which comes from a lot of different sources, stock exchange applications need to integrate data from different sources and to offer different facets for those data, organizational portals are inherently needing web data for their organizational users and also need to structure data for their own uses, and advertising need various semantically descriptions for the products. It seems that the main issues involved by the relatively small application of the semantic web technology in the business domain relates to technical aspects that semantic web technology poses and not to the business domain. In contrary, the business domain seems to require this promising semantic web technology.

The current promises of semantic web technologies are for 1) data integration by interposing a meta data layer which semantically describes data and for 2) semantic search by offering end-users different web sources as response to the key words that they use in their web searches.

The standard format to represent data for semantic web applications is Resource Description Format (RDF) which will be treated by this article in a following section.

We address the problem of scalability for semantic web business application by discussing the results obtained from a survey through which we intended to define the necessary variables in assuring the scalability. In the next section we discuss the research model proposed by the present article. Another section is dedicated to survey's analyze and another one to the scalability model. Discussions and conclusions are presented in different sections.

## 2 Problem Formulation

We initiate the research by discussing the problem statement. It seems that, for the moment, there is a whole theory related to semantic web, there is available a set of Internet technologies, but it is not quite clear what are the best semantic web technologies needed to assure the scalability of semantic web applications.

Particularly, an algorithm, design, networking protocol, program, or other system is said to scale if it is suitably efficient and practical when applied to large situations (e.g. a large input data set, a large number of outputs or users, or a large number of participating nodes in the case of a distributed system).

Normally the semantic model would have the capacity to scale, because the semantic model is designed to create links between various meanings. It is normal to think that all these networks of meanings will create technical problems, but also applicability problems.

The studies dedicated to scalability discuss the horizontal scalability and vertical scalability. To scale horizontally (or scale out) means to add more nodes to a system, such as adding a new computer to a distributed software application. An example might be scaling out from one Web server system to three. To scale vertically (or scale up) means to add resources to a single node in a system, typically involving the addition of CPUs or memory to a single computer. Such vertical scaling of existing systems also enables them to use virtualization technology more effectively, as it provides more resources for the hosted set of operating system and application modules to share.

When discussing the semantic web scalability we must cite these sources:

- Prolog-based Infrastructure for RDF: Scalability and Performance – storage has achieved about 40 million triples [9]
- 3store: Efficient Bulk RDF Storage - has achieved about 20 million triples and 5000 classes and properties [5]
- Tucana Semantic Analysis and Interoperability – has achieved about 100 million statements (32 bit) or 1 billion statements (64 bit), with most recently the 32 bit increased to 350 million RDF statements (Northrop Grumman on Tucana) [8]

We will treat the problem of scalability from the semantic model point of view and from the semantic web technology point of view. In order to do this we proposed a research model, we defined some variables in order to test the research hypothesis that we proposed and, finally, we defined a scalability model. We treat in the next section the research design.

## 3 Research design

There have been many studies on the systems' scalability. The current semantic web technology have to offer scalability in order to have success.
There are some questions which arise in this context: 1) in which way the semantic web model can raise the scalability of actual enterprise systems? 2) what are the particularities of semantic web technologies which can create problems to the scalability of semantic web model?

Most of all previous studies focus on technical scalability [7, 5].

In this study, we focus on analyzing the technological limitations which are interrelated with the requirements of semantic web applications in

business domain in order to delimitate a scalability model. We propose the research model shown in Fig. 1.

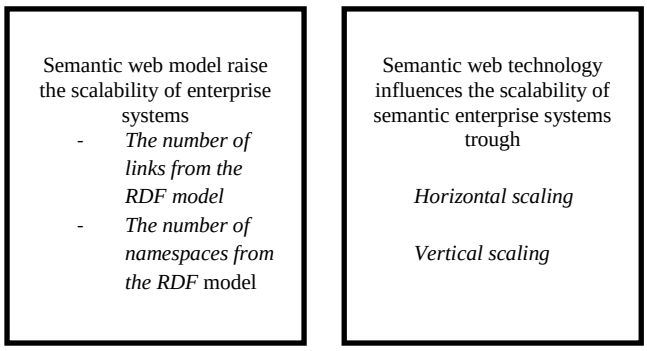


Fig. 1 The research model

Table 1 provides the definition of the factors listed in Fig. 1, and indicates the number of items used to measure each construct.

Table 1. Definition of the variables considered by the scalability model of business semantic web application

| Factors/ Variables | Definition |
|---|---|
| 1.Factors determined by the semantic web model | |
| 1.1 The number of URIs | The number of individual entities which have attached semantic meanings |
| 1.2 The number of namespaces | The number of ontologies/ vocabularies |
| 2. Factors determined by semantic web technology | |
| 2.1 Vertical scaling | Improving the hardware performance on the same machine |
| 2.2 Horizontal scaling | Semantic web business application processed on multiple machines (farm servers) |

We derived two main hypotheses from the research model.

H1: the semantic web model presents advantages for scaling enterprise systems to bigger data sets and to offer more meanings for information

H2: the semantic web technology influences the technical scalability of semantic web business applications.

# 4 The scalability model

In the following, we divide our discussion into different aspects to treat according to our research design: semantic web technology and semantic model.

## 4.1. Semantic Web technology

Semantic web found its practical applications in the form of Linked Open Data. The goal of the W3C Semantic Web Education and Outreach group's Linking Open Data community project is to extend the Web with a data commons by publishing various open datasets as RDF on the Web and by setting RDF links between data items from different data sources. In October 2007, datasets consisted of over two billion RDF triples, which were interlinked by over two million RDF links [4]. By September 2011 this had grown to 31 billion RDF triples, interlinked by around 504 million RDF links.

Tim Berners-Lee outlined four principles of linked data in his Design Issues: Linked Data note, paraphrased along the following lines:[2]

1. Use URIs to identify things.
2. Use HTTP URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.
3. Provide useful information about the thing when its URI is dereferenced, using standard formats such as RDF/XML.

Include links to other, related URIs in the exposed data to improve discovery of other related information on the Web.

Vertical scaling consists in replacing existent hardware resources with new ones with a higher degree of performance.

Semantic web needs high processing speed and for this not only the processing speed of the CPU us useful but the speed to access data is also very important.

Therefore there is a need that the semantic database be available on memory storages which offer fast data access.

For this, from the existent memory storages the most suitable to assure fast data access is the Random Access Memory (RAM). The most common RAM capacity existent on the market is of 8 GB. This means that the models which can be loaded by this dispositive cannot be bigger than 8 GB. We observe a limitation on the performance of semantic database.

Given this limitation any semantic web application developer will look to assure the horizontal scalability.

The horizontal scalability raises the problem of partitioning a big data set on multiple systems which can be accessed in parallel. In order to realize this there is a need of a data management system which can assure the sharing of files so that any request on data be routed directly to the system which has the most relevant data to answer to the queries.

But the horizontal scaling not only requires a semantic data management system. It is also a problem of partitioning namespaces because every URI is described by different namespaces. We will discuss the semantic model in the next subsection to understand better the problem of semantic relations.

Open Data generates economic benefit especially when data can be linked in new ways according to an online surveys conducted by SWC in March 2012. 75,7% from 113 respondents agreed with this fact. As for the most relevant options for enterprises to make use of open data the respondents said that these are: applications (73,9%), business intelligence (63%), vizualization (71,7%) [3].

There are a lot of data sets which form together the Cloud Diagram. We ennumerate: Dbpedia, Freebase, OpenCyc, Geonames, LinkedGeoData, BBC, data.gov.uk and data.gov, DBLP, UniProt, KEGG, PubMed, Gene Ontology.

Considering the problem of existing vocabularies for retail and commerce we discuss the GoodRelations ontology ( http://purl.org/goodrelations/ http://www.bestbuy.com/) has provided a richer ontology for describing many aspects of e-commerce, such as businesses, products and services, offerings, opening hours, and prices. GoodRelations has seen significant uptake from retailers such as Best Buy and Overstock.com (http://www.overstock.com/ ) seeking to increase their visibility in search engines such as Yahoo! and Google, that recognise data published in RDFa using certain vocabularies and use this data to enhance search results.

Building semantic web applications means to integrate data from different sources described with different vocabularies. The most often used vocabularies are: Friend-of-a-Friend (FOAF), SIOC, DOAP, Dublin Core, Review Vocabulary, GoodRelations, Music Ontology, Organization Ontology, Linking Open Description of Events (LODE). Google, Yahoo! and Bing are using schema.org which is a vocabulary that can be integrated directely into html pages. The operators of the world's largest search engines propose to mark up website content as metadata about itself, using microdata, according to their schemas. Those schemas can be recognized by search engine spiders and other parsers, thus gaining access to the meaning of the sites.

Currently the main problems that semantic web technology presents for the practical application are: 1) the relative exponential growth of the datasets, namespaces, SPARQL endpoints, 2) the lack of proper semantic web technologies useful for managing RDF triples in a distributed manner and 3) the lack of understanding the potential of this technology for the web sites owners. The whole concept of Open Linked Data consists in publishing and consuming linked open data. Even if technologies are available the main step is represented by publishing semantic data and this step can be realized only by web sites owners.

The fault has been in the software paradigm and indexing paradigm and perhaps other data representation paradigms, because be it an AI language, existing RDBMS or OODMS data system, or even a “native” storage system, they all are limited by 250 million triples, and most choke well before that. While useful, we should not have to look to 64 bit computing to overcome these limits, but rather combine it with a better understanding of what is appearing to be a truly unique data storage paradigm.

## 4.2. Semantic model

The Resource Description Format (RDF) is a data model usefull to describe resources as subject-predicate-object. There are two serialization formats: RDF/XML and RDFa. The subject is an URI which describes the resource. The predicate is identified by an URI which belongs to an ontology. The object is a literal or an URI to another resource related to the subject.

Every vocabulary/ ontology has a namespace which can be addressed locally if the semantic web application uses a triple store or on the web. An example of a locally stored vocabulary is presented in Fig. 2. Usually a vocabulary contains classes and properties.

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:swrc="http://ontoware.org/swrc/swrc_v0.3.owl#"
xml:base="http://www.example.org/researcher#">
<rdf:Description rdf:ID="AcademicEvent">
<rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
</rdf:Description>
<rdf:Description rdf:ID="Conference">
<rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
<rdfs:subClassOf rdf:resource="http://ontoware.org/swrc/swrc_v0.3.owl#Conference"/>
<rdfs:subClassOf rdf:resource="#AcademicEvent"/>
</rdf:Description>
```

Fig. 2. The researcher ontology

In the linked data, usually every class which belongs to a vocabulary is described as a sub-class of at least another vocabulary. Also, every class is described by using properties belonging to different vocabularies. Once the ontology is designed and available anyone can describe data according to the ontology.

Given the above examples it is easy to imagine that the management of the relations created by referring to various vocabularies and URIs is difficult to partition in order to assure the horizontal scaling.

After discussing the aspects related to the semantic model and to the semantic web technology and basing on our variables we provide the architecture elements that respect or scalability model.

The architectural elements needed for the vertical scaling (possible up to the limit of 8 GB of one RDF file) are the quality of the semantic model provided by the semantic web application developer (the number of URIs, the number of namespaces).

The architectural elements needed for the horizontal scaling (needs the existence of distributed management systems) – are the quality of the semantic descriptions realized by every site owner and the availability of proper vocabularies in order to describe data.

## 5 Discussions

In this paper we conducted a methodology which intended to find the variables needed in order to define the scalability of the business semantic web applications. We wanted to demonstrate that the users of computer-based applications from business activities are using information that comes from different sources. We also observed the fact that these users consider that they use a lot of digitalized information in their activities but this information is not properly integrated.

Starting from the existing studies and theories we identified the main variables that our survey addresses in order to propose the architectural elements needed for semantic web based applications.

We found out that the semantic model has the ability to scale enterprise computer-based applications by its capacity to model relations between concepts, instances and properties.

We also found out that users need and want this kind of semantic interoperability.

In the end, we discussed the main challenges that the existent semantic web technologies have for the proper application in the business domain and we found out that even if there are available the necessary hardware capacities there still are some problems to address. These problems relate to the necessary technology needed to work with the semantic model.

We also found out that there are differences between the semantic web application developer, the end-user and the site owner. Where the semantic web application developer overcome the challenge of managing diverse RDF files, there is a definitely necessity that every site owner describe its data according to common and accepted vocabularies developed by the semantic web community.

## 6 Conclusion

The main contribution of our paper consists in defining the architectural aspects of the semantic web business applications which raise problems in assuring the scalability.

We consider that our model can be a framework to analyze the limits and the advantages of current semantic web applications and also a framework for the aspects important to take into consideration in semantic web applications developing.

Given the practical usefulness we expect that actual semantic web based developers realize that it is not possible for the moment to scale the semantic web technology to the entire Internet sets of data. Even so, with careful considerations valuable semantic web applications can be developed. There are not impossible limits at all. Of course there are limits for semantic Google but not for semantic web applications in business.

Our model has implications for software engineering given the aspects concerning the semantic model. Also, we consider that our model has implications in establishing good semantic web application requirements.

## Acknowledgement

This work was supported by CNCSIS-UEFISCSU, project number PN II-RU code 188/2010.

*References:*

[1] T. Berners Lee, 2000, http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

[2] T. Berners-Lee, T., 2006, Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html

[3] A. Blumauer, SWC, 2012, Open Data for Enterprises, http://www.slideshare.net/ABLVienna/open-data-for-enterprises-12126124

[4] D. Fensel, F. M. Facca, E. Paslaru Bontas Simperl, I. Toma: Semantic Web Services. Springer 2011: I-XI, 1-357

[5] S. Harris, N. Gibbins, 3store: Efficient Bulk RDF Storage, Project Report,

http://km.aifb.kit.edu/ws/psss03/proceedings/harris-et-al.pdf, accessed August 2012
[6] I. Herman, W3C Semantic Web Activity, W3C. http://www.w3.org/2001/sw/, Retrieved March 13, 2008.
[7] R. Lee, MIT, 2004 http://simile.mit.edu/reports/stores/, accessed August 2012
[8] Tucana Technologies INC, 2004, Massive Scalability for RDF Storage and Analysis, available at www.w3c.org
[9] J. Wielemaker, G. Schreiber, B. Wieling, Prolog-based Infrastructure for RDF: Scalability and Performance, Project Report, http://www.cs.vu.nl/~guus/papers/Wielemaker03a.pdf, acessed August 2012
[10] World Wide Web Consotium, www.w3c.org
[11] Linked Open Data, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData, http://linkeddata.org/
[12] Semantic Web Use Cases, http://www.w3.org/2001/sw/sweo/public/UseCases/
[13] RDB2RDF Incubator, http://www.w3.org/2005/Incubator/rdb2rdf/
[14] D2RQ Platform, http://www4.wiwiss.fu-berlin.de/bizer/D2RQ/spec/
[15] OpenLink Virtuoso Platform, Automated Generation of RDF Views over Relational Data Sources, http://docs.openlinksw.com/virtuoso/rdfrdfviewgnr.html
[16] Securities Exchange Commission RDF Data, http://www.rdfabout.com/demo/sec/
[17] W3C Semantic Web Frequently Asked Questions. W3C. http://www.w3.org/2001/sw/SW-FAQ. Retrieved March 13, 2008.